Bioinformatics, 31(23), 2015, 3815–3821
doi: 10.1093/bioinformatics/btv455
Advance Access Publication Date: 7 August 2015
Original Paper

Data and text mining

Impact of normalization methods on high-throughput screening data with high hit rates and drug testing with dose–response data

John-Patrick Mpindi*, Potdar Swapnil, Bychkov Dmitrii, Saarela Jani, Khalid Saeed, Krister Wennerberg, Tero Aittokallio, Päivi Östling and Olli Kallioniemi

University of Helsinki, Institute for Molecular Medicine, Tukholmankatu 8, FI-00290, Helsinki, Finland

*To whom correspondence should be addressed.
Associate Editor: Jonathan Wren

Received on April 27, 2015; revised on July 10, 2015; accepted on July 30, 2015

Abstract

Motivation: Most data analysis tools for high-throughput screening (HTS) seek to uncover interesting hits for further analysis. They typically assume a low hit rate per plate. Hit rates can be dramatically higher in secondary screening, RNAi screening and in drug sensitivity testing using biologically active drugs. In particular, drug sensitivity testing on primary cells is often based on dose–response experiments, which pose a more stringent requirement for data quality and for intra- and inter-plate variation. Here, we compared common plate normalization and noise-reduction methods, including the *B*-score and the Loess a local polynomial fit method under high hit-rate scenarios of drug sensitivity testing. We generated simulated 384-well plate HTS datasets, each with 71 plates having a range of 20 (5%) to 160 (42%) hits per plate, with controls placed either at the edge of the plates or in a scattered configuration.
Results: We identified 20% (77/384) as the critical hit-rate after which the normalizations started to perform poorly. Results from real drug testing experiments supported this estimation. In particular, the *B*-score resulted in incorrect normalization of high hit-rate plates, leading to poor data quality, which could be attributed to its dependency on the median polish algorithm. We conclude that a combination of a scattered layout of controls per plate and normalization using a polynomial least squares fit method, such as Loess helps to reduce column, row and edge effects in HTS experiments with high hit-rates and is optimal for generating accurate dose–response curves.
Contact: john.mpindi@helsinki.fi
Availability and implementation, Supplementary information: R code and Supplementary data are available at *Bioinformatics* online.

1 Introduction

High-throughput drug testing is increasingly being applied on, e.g. established cancer cell lines, drug-resistant cancer cell models, primary cancer cells, iPS and other stem cell models of disease (Barretina *et al*., 2012, Crystal *et al*., 2014, Gao *et al*., 2014, Pemovska *et al*., 2013, Shoemaker, 2006, Tyner *et al*., 2013). This facilitates investigation of the functional effect of a spectrum of drugs on representative cell models and may be developed as a tool for cancer diagnostics and personalized medicine in the future (Barretina *et al*., 2012, Garnett *et al*., 2012, Pemovska *et al*., 2013, Shoemaker, 2006, Tyner *et al*., 2013, Yang *et al*., 2013). For example, we have developed drug sensitivity and resistance testing (Pemovska *et al*., 2013) for primary *ex-vivo* cancer cells from leukemia patients using serial dilutions of a comprehensive drug panel, previously containing 187, but now 461 preclinical and clinical cancer drugs (Pemovska *et al*., 2013). As opposed to high-throughput

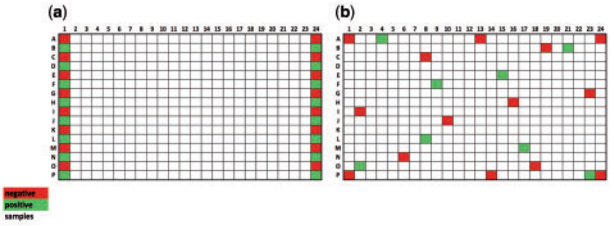

Fig. 1. Plate layouts commonly used in drug testing experiments. (**a**) Layout based on placing controls in column 1 and 24. (**b**) Layout based on scattering controls across the entire plate

screening (HTS) with a discovery intent, the drug testing is done with known bioactive agents, often leading to high hit rates per plate. Also, there is a need to generate dose–response curves for the quantitative assessment and validation of the results. Hence, the quality requirements for primary data are more stringent than in regular HTS experiments. Primary cells from patients will most often not be available for replicate HTS experiments or validation screens. Recent articles have raised concerns about the lack of reproducibility of drug testing data even in established cell lines and have suggested standardization of laboratory protocols, drug annotations and drug-response scoring metrics (Hatzis *et al*., 2014, Yadav *et al*., 2014). There is little data on the performance of normalization methods for HTS experiments where hit rates are commonly high.

Like all HTS experiments, drug-testing studies rely on controls to normalize data points within and between plates. We then generate percent inhibition metrics and produce dose–response curves based on serial drug dilutions. The way these dilutions are done and how controls are applied on a plate layout often varies. For example, it is technically easier to make a plate with the highest concentration of each drug and then make serial dilutions across the entire plate. This will result in plates with distinctly different overall hit rates. Second, a plate layout can be designed to include positive and negative controls in the first or last column or in random positions (scattered layout, Fig. 1). If controls are placed in the first or last column, they are sensitive to edge effects, which often occur due to evaporation. While scattered layout would be optimal, the layout is often chosen based on what is technically most feasible.

High-throughput experiments manifest systematic row, column and edge effects when a global signal distribution surface is analyzed (Makarenkov *et al*., 2007). Thus, there is a need for normalization and analysis methods that reduce false positives in the HTS experiments (Dragiev *et al*., 2012, Kwan and Birmingham, 2010, Murie *et al*., 2013, Rieber *et al*., 2009, Seiler *et al*., 2008). For instance, a recent study (Murie *et al*., 2013) suggested the use of control plate regression method, which depends on adjusting signal intensities on the treatment plates by scaling the data based on bias estimates on a control plate.

This approach requires an extra control plate, and it is impossible to calculate quality control (QC) metrics per plate without having both the positive and negative controls on each of the treatment plates. Moreover, dispensing of reagents may introduce plate-specific variations stressing the importance of having controls on each plate. Another approach would be to adapt methods developed for normalizing microarray experiments such as generalized procrustes analysis (Xiong *et al*., 2008) and modified Loess [loessM (Risso *et al*., 2009)]. Generalized procrustes analysis requires replicate experiments, whereas loessM requires an experimental design based on biological replicates dye-swap. Because of the limited amount of cells available, replicate HTS experiments cannot often be done with primary cells and hence methods that depend on replicates can be difficult to implement.

The *B*-score (Brideau *et al*., 2003, Liu *et al*., 2013, Malo *et al*., 2006) is perhaps the most popular normalization method for HTS experiments. Importantly, however, the *B*-score assumes a low hit rate and is based on the iterative application of the Tukey median polish algorithm. In addition to the *B*-score, a number of other noise reducing methods based on polynomial least squares fit have been published (Makarenkov *et al*., 2007) and implemented in open source Bioconductor packages cellHTS (Boutros *et al*., 2006) and RNAither (Rieber *et al*., 2009) and platforms such as HitPick (Liu *et al*., 2013). For example, the locally weighted scatterplot smoothing (Loess-fit) is based on fitting a distribution surface on the whole plate data matrix. Despite the value of performing data normalization on datasets affected by within-plate effects, normalization methods often introduce bias when applied on any dataset (Dragiev *et al*., 2011). Hence, it is necessary to assess whether data from an HTS experiment are compatible with the original assumptions of the normalization methods. Here, we will evaluate the reproducibility and quality of simulated and real data normalized with the *B*-score and the Loess-fit approaches.

2 Methods

2.1 Datasets

To compare the reproducibility and quality of drug testing data after normalization for row, column and edge effects within a single plate, we utilized a simulated dataset and a real experimental dataset from the FIMM-High Throughput Biomedicine facility. The experimental data consisted of testing the effects of 306 FDA approved and investigational drugs on the viability of two prostate cancer cell lines VCaP and LAPC4. The screens were performed in replicate, and each drug was screened twice across five concentrations (Fig. 2b). The drug-testing pipeline is described in detail previously (Pemovska *et al*., 2013, Yadav *et al*., 2014). The plate layout is given in Supplementary Material (Supplementary File S2), including the position of controls. The simulated dataset consisted of 142 plates with each plate designed to contain 306 drugs and well controls mimicking the distribution pattern found in real HTS data. We mimicked the distribution of real data to make the QC metrics to resemble those expected in real data. We tested the normality of the positive and negative control distributions in real data using the Shapiro–Wilks (Shapiro and Wilk 1965) method from R *stats* package. The test showed that the real data follows a normal distribution ($W = 0.9916$, P value $= 0.4696$ for negative controls and $W = 0.9725$, P value $= 0.2197$ for positive controls). We also further confirmed the assumption of normality using an external HTS dataset of two duplicate 384-well plates containing dimethyl sulfoxide (DMSO) negative controls and no treatment ($W = 0.9978$, P value $= 0.9007$ and $W = 0.9961$, P value $= 0.4602$). The data were downloaded from CHEMBANK (http://chembank.broadinstitute.org/assays/view-project.htm?id=1001118) by selecting CellTiterGlo(1135.0009).

We generated an increasing number of hits on each of the 71 plates up to a hit rate of 42% (160 drugs). In real drug sensitivity testing experiment, the number of hits keeps increasing with increasing drug concentration level. Therefore, in our simulation experiment, we increased the number of hits iteratively by adding 2 hits on each run starting from 5% (20) hits until a hit rate of 42% (160 drugs). We started by producing data for 287 drugs considered as non-hits sampled from a distribution of negative controls $N(\mu_1, \delta_1)$ and added 20 drugs considered as hits sampled from a

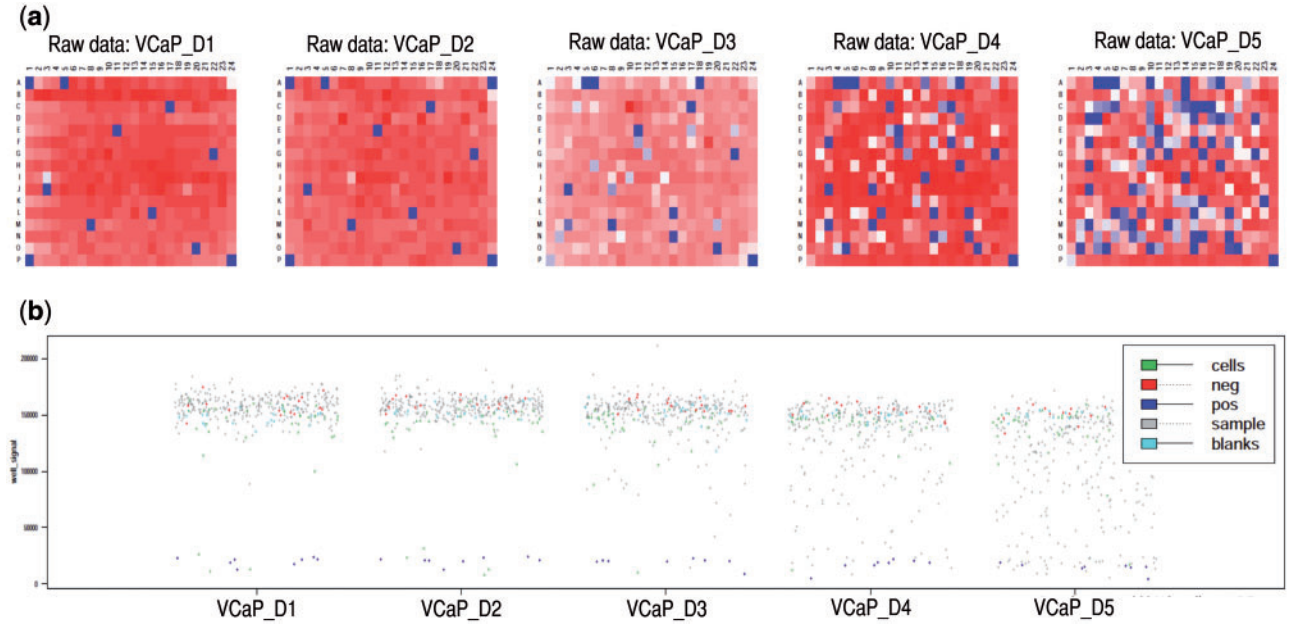

Fig. 2. Data quality visualization by heatmaps and scatter plots (**a**) Comparing quality of data per drug testing plate starting from the plate containing the lowest concentration of drugs (D1-left) to that containing highest concentration drugs (D5-right). (**b**) The scatter plot shows that the positive (blue) and negative (red) controls clearly separate in all the five plates, which indicates good screen quality

distribution of positive controls $N(\mu_2, \delta_2)$. For the remaining 70 plates, we reduced the non-hits by 2 and increased the number of hits by the same number on each run. These steps helped to generate a drug testing dataset with an increasing number of hits per plate based on two plate layouts.

2.2 Data preprocessing

Data from all simulated and real drug testing experiments were preprocessed using R statistical software. Data outputs from the Pherastar FS plate reader (Ortenberg, Germany) were converted to matrices and visualized as heatmaps and well scatters. Pre- and post-normalization QC assessment was performed using *Z*'-factor (Zhang *et al.*, 1999) and strictly standardized mean difference (SSMD) (Zhang, 2007).

2.3 Quality control

We performed QC on both the pre- and post-normalization data. Here, we implemented the commonly used QC methods including SSMD and *Z*'-factor. The *Z*'-factor was chosen because the calculation is based on using controls which are essential for calculating percent inhibition.

$$Z' - \text{factor} = 1 - 3(\delta_{\text{h.c}} + \delta_{\text{l.c}})/\mu_{\text{h.c}} + \mu_{\text{l.c}} \quad (1)$$

where h.c = high control represents the signal detected from negative control wells (DMSO) leading to no effect on the cells (no decrease in viability), while l.c = low control refers to signal intensities from positive control wells (benzethonium chloride) leading to the greatest cell killing effect (maximal decrease in viability). The *Z*'-factor values can range from negative infinity to one, where values > 0.5 represent a very good experiment, >0 and <0.5 as moderately good experiment and <0 as a poor experiment. We also implemented the SSMD statistic as it offers a robust assessment of the quality of the screen. SSMD has been shown to be more accurate and less conservative indicator of quality than the *Z*'-factor (Birmingham *et al.*, 2009). In our work, we also preferred results from SSMD since it gives a more robust indication about the quality of data from control wells.

$$\text{SSMD} = \mu_{\text{h.c}} - \mu_{\text{l.c}} / \sqrt{(\delta^2_{\text{h.c}} + \delta^2_{\text{l.c}})} \quad (2)$$

where h.c = high control and l.c = low control are used as described above.

2.4 Percent inhibition

To be able to compare data analyzed on different plates across several concentrations, we calculated the percent inhibition metric. Percent inhibition values greatly depend on the quality of data acquired from control wells. Controls are often placed at the edges of the plate (first and last column) where they are impacted by edge effects. Here, we compared this with a scattered one. Given data on a plate p, we calculated the percent inhibition (I_{ijp}) for the value at row i and column j as follows:

$$I_{ijp} = (\mu_{\text{h.c}} - x_{ij}/\mu_{\text{h.c}} - \mu_{\text{l.c}}) * 100 \quad (3)$$

where x_{ij} is the measured signal value at row i and column j on the pth plate, $\mu_{\text{h.c}}$ is the mean of negative control sample value on plate p and $\mu_{\text{l.c}}$ is the mean of the positive control sample values on plate p.

2.5 Within-plate normalization and corrections

Systematic errors due to, e.g. row, column and edge effects occur in drug testing experiments and can significantly affect the downstream analysis of drug testing data. Using standard QC methods and whole plate visualization methods, it is possible to determine whether within-plate noise corrections are needed, although often this is performed automatically.

The percent inhibition calculation takes care of the cross-plate systematic effects by normalizing all values to a percent score, whereas within-plate effects cannot be corrected. A common approach for correcting row, column and edge effects is the *B*-score.

B-score calculation is robust to outliers since it utilizes a non-parametric approach based on Tukey's median polish algorithm (Kafadar, 2003). Given plate p, where x_{ij} is the measured signal value at row i and column j, we calculate the *B*-score as follows:

$$B\text{-score}_{ijp} = \text{med.polish.fit.sample.residual}(r_{ijp})/\text{MAD}_p \quad (4)$$

where r_{ijp} represents the two-way fitted median polish residuals calculated iteratively to minimize row and column effects using the medpolish function in the stats package of R software. MAD for plate p refers to the median absolute deviation calculated from the r_{ijp} values as follows:

$$\text{MAD}_p = \text{median}\{|r_{ijp} - \text{median}(r_{ijp})|\} \quad (5)$$

The *B*-score method assumes a low hit rate on the row and column, which does not hold for drug testing data in particular for plates containing drugs applied at high concentration. To address this concern, we tested an approach based on fitting a local distribution surface using least squares polynomial approximation (Makarenkov *et al.*, 2007). We performed local regression on a single plate using the Loess-fit method by assessing the deviation of each fitted value from the median. Extreme deviations of data from locally adjacent wells would suggest the existence of systematic within-plate errors causing peaks and valley shapes in the smooth surface fit. A well correction is then performed by subtracting from or adding to the original value. Given plate p, where x_{ij} is the measured signal value at row i and column j, we calculated the loess-fit result $\hat{x}_{ij}$ as follows.

$$\hat{x}_{ij} = x_{ij} - (\text{loess.fit} - \text{median}(\text{loess.fit}_{ij})) \quad (6)$$

where loess.fit_{ij} is the value from loess smoothed data at row i and column j calculated using the loess function in stats package of R software with a span of 1. The current implementation of loess assumes (i) that the controls are scattered across the plate and (ii) that drug hits on the plate are randomly distributed. The R code for performing loess normalization is provided under Supplementary Material (Supplementary File S3). Next, we performed the cross-plate normalization using the percent inhibition formula above. Then, we used the percent inhibition values to examine the reproducibility of the post-normalization data using the reproducibility concordance correlation coefficient (rccc) (Lin, 1989) implemented in the epiR package of R software.

3 Results and discussion

3.1 Visualization and QC of high hit-rate and dose–response experimental data

To demonstrate the importance of raw data visualization and QC steps, we analyzed our in-house drug testing dataset of two prostate cancer cell lines screened in replicate. Each of the replicate screens contained five 384-well plates seeded with cells and incubated with 306 drugs and controls. The controls for Cell Titer-Glo (CTG) viability assay included 16 negative controls with DMSO only and 8 positive control wells with 100 μM benzethonium chloride. In addition, 19 of the remaining wells were left blank and 35 wells contained cells only. The drugs were plated in five different concentrations in 10-fold dilutions covering a 10 000-fold concentration range. First, we visualized the 384-well plate raw signal intensities as a heatmap (Fig. 2a) to show the distribution of the high (red) and low (blue) hits. The heatmap visualization helps to detect systematic errors due to, e.g. cell seeding (stripping, checkerboard) or evaporation (edge-effects). The heatmaps were arranged according to increasing dose of each drug (D1–D5) (Mangat *et al.*, 2014).

Table 1. Pre- and post-normalization QC scores

Plate_ID	Raw_ zprime	*B*-score_ zprime	Loess-fit_ zprime	Raw_ ssmd	*B*-score_ ssmd	Loess-fit_ ssmd
LAPC4_D1_rep1	0.63	0.67	0.64	10	12	11
LAPC4_D2_rep1	0.65	0.62	0.67	11	10	12
LAPC4_D3_rep1	0.64	0.66	0.73	11	12	15
LAPC4_D4_rep1	0.63	0.66	0.65	10	12	11
LAPC4_D5_rep1	0.59	**0.48**	0.51	9	8	8
LAPC4_D1_rep2	**0.46**	**0.41**	**0.44**	**6**	**6**	7
LAPC4_D2_rep2	0.57	0.58	0.61	8	9	10
LAPC4_D3_rep2	0.61	0.63	0.65	10	10	11
LAPC4_D4_rep2	0.54	0.69	0.61	8	13	10
LAPC4_D5_rep2	0.55	**0.36**	0.55	8	**6**	8
VCap_D1_rep1	0.75	0.78	0.73	15	19	15
VCap_D2_rep1	0.82	0.81	0.75	22	20	15
VCap_D3_rep1	0.8	0.8	0.82	20	20	22
VCap_D4_rep1	0.76	0.78	0.79	17	18	20
VCap_D5_rep1	0.73	**0.4**	0.61	14	**6**	10
VCap_D1_rep2	**0.05**	**0.07**	**−0.03**	**4**	**4**	**3**
VCap_D2_rep2	**0**	**−0.01**	**−0.07**	**3**	**3**	**3**
VCap_D3_rep2	**−0.77**	**−0.7**	**−0.78**	**2**	**2**	**2**
VCap_D4_rep2	**−0.61**	**−1.13**	**−0.82**	**2**	**2**	**2**
VCap_D5_rep2	**−0.18**	**−0.88**	**−0.39**	**3**	**2**	**3**

Column headings show the QC score used per data type. The quality of each screen was assessed pre- (Raw) and post-normalization (after normalization) by either *B*-score or Loess-fit. Low QC scores below the recommended threshold are highlighted in bold.

As expected, the number of hits increased with the increasing drug dose. Second, plate-well scatters were used to illustrate the overall quality and reproducibility of the HTS experiment based on examining the performance of control wells across the five drug dose levels (Fig. 2b) arranged in ascending order. As can be seen, plates containing drugs applied at low dose (D1–D2) contain fewer outliers or hits compared to plates with higher doses of drugs (D3–D5). High or low signal values at the edge of a plate highlight edge effects.

To automatically flag single plates with artifacts that need correction, we employed a quantitative approach based on calculation of the *Z*'-factor and the SSMD scores (*Z*' and SSMD, Table 1) (Zhang, 2011; Zhang *et al.*, 2007). All plates with an SSMD score less than 6 were flagged for visual inspection and further correction. *Z*'-factor and SSMD QC metrics were able to highlight plates likely to contain row, column and edge effects. It is important to visualize the data using plate heatmaps and plate-well scatters since the metrics tend to be skewed by the presence of outliers among the control samples.

With visual data quality inspection, one can detect the presence of outliers among controls placed on each plate. For our experiments, we observed that when the QC metrics were good, there was always a good correlation between the replicate experiments. The X–Y correlation plot in Figure 3a illustrates the effect of within-plate effects on the data reproducibility for two replicate VCaP cell line screens leading to low rccc of 0.82 (confidence interval: 0.80, 0.83).

The discordant results were obtained for the VCaP cell line screen replicate 2 with poor QC values (Table 1, Fig. 3a) compared to a LAPC4 cell line screen with better quality raw data resulting in

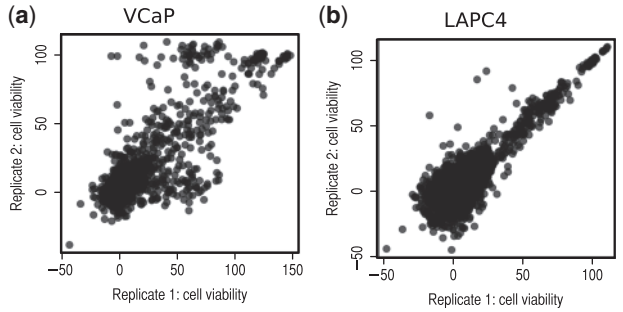

Fig. 3. Reproducibility of un-normalized data from replicate experiments. The scatterplot shows the correlation of raw data from two replicate drug testing screens for the VCaP and LAPC4 cell lines. The global rccc score was calculated by putting together percent inhibition values for the five plates to make one plot and rccc score. The VCaP (**a**) drug testing experiment showed lower reproducibility due to the poor quality of replicate experiment 2 thus leading to an rccc score of 0.82 (confidence interval: 0.80, 0.83) compared to LAPC4 (**b**) rccc score of 0.90 (confidence interval: 0.89, 0.90)

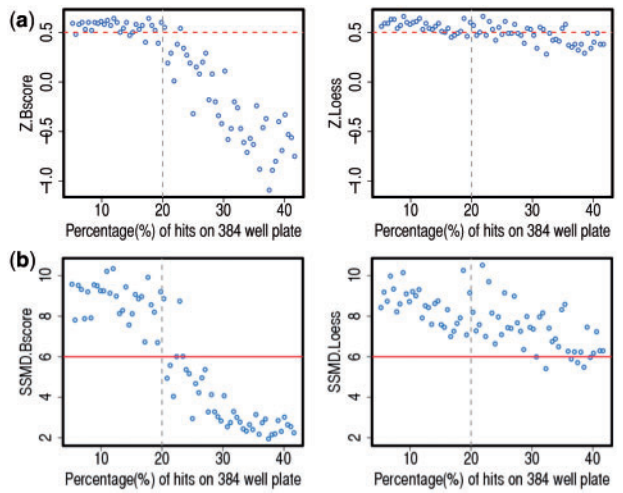

Fig. 4. Identifying the maximum tolerable hit rate needed to perform normalization. QC estimates (**a**) *Z*′-factor and (**b**) SSMD as a function of the percentage (%) of hit-drugs (drugs showing high cell killing pattern similar to the positive controls placed on the same plate). The maximum tolerable hit-rate on any plate was identified as 20% indicated by the dotted grey line. The horizontal dotted red line for *Z*′-factor and the solid line for SSMD indicate the recommended QC threshold for a good screen

rccc score of 0.90 (confidence interval: 0.89, 0.90, Table 1, Fig. 3b). These findings indicate that poor QC scores caused by within-plate systematic errors introduced false-positive hits (outliers) in the VCaP experiment. Drug plates with poor QC scores can lead to inaccurate percent inhibition calculations needed for calculation of dose–response metrics and curve fitting.

Since within-plate effects can lead to low reproducibility and poor QC results, we tested whether normalization algorithms could be used to improve the quality and reproducibility of the data. Since the within-plate normalizations were not designed for high hit rate scenarios, we wanted to specifically test whether they could result into post-normalization QC scores that are lower than those obtained from the raw data.

3.2 Simulation of QC metrics under increasing hit rate

We applied two simulations mimicking the commonly used plate layouts under increasing hit rate scenarios to systematically identify

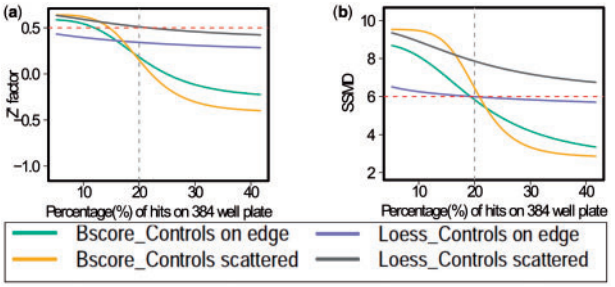

Fig. 5. Simulated data results showing increasing hit rate versus post-normalization QC scores. QC estimates based on *B*-score and Loess normalized data (**a**) *Z*′-factor and (**b**) SSMD. The curves show the change in QC scores as the hit rate increases. Each normalization method was tested on two plate layouts outlined on the figure legend (Controls on edge and Controls scattered)

the ideal hit rate and plate layout combination that would enable normalization methods correcting for within plate effects. Our simulation study consisted of 142 plates, each plate comprising 384 wells seeded with 306 drugs, 16 negative controls and 8 positive controls. The controls were placed either on the edge or randomly scattered across the entire plate comprising a total of 71 plates with each type of control layout.

First, we generated data for each of the 142 plates by mimicking an increasing hit rate often experienced in drug testing experiments. Hits and non-hits were generated from two independent normal distributions $N(\mu_1, \delta_1)$ and $N(\mu_2, \delta_2)$ mimicking realistic signal intensities with non-hits being generated from a distribution close to negative control wells signal ($\mu_1 = 122\,674.6$, $\delta_1 = 10\,481.78$), while the hits were generated from a distribution close to the positive control wells signal ($\mu_2 = 38\,158.6$, $\delta_2 = 10\,481.78$). We started with a hit rate of 20 wells out of 384 wells (5%) and increased it iteratively by adding 2 hits on each new plate until the cumulative hit rate was 160 wells (42%). We next examined the effect of normalizations on the quality of the data by inspecting *Z*′-factor and SSMD QC metrics for each of the 142 plates. After generating post-normalization *Z*′-factor and SSMD QC scores (Supplementary File S1), the critical point was determined to be when the QC scores dropped below the recommended thresholds (*Z*′-factor < 0.5 and SSMD < 6). The maximum tolerable hit rate to carry out normalizations was 20% or 77/384 wells. Beyond this level, normalizations started to affect the data quality severely as illustrated in Figure 4 on plates with a scattered layout. The mean *Z*′-factor score based on Loess-fit normalized data was 0.5 for a hit rate window between 5% and 42% compared with the mean *Z*′-factor score of 0.08 based on *B*-score normalized data generated from plates with the scattered layout for controls. We then examined whether post-normalization QC scores were significantly different between the two plate layout formats.

The results in Figure 5 indicate post-normalization QC scores represented as a function of an increasing hit rate. Consistent with the observations in Figure 4, the *B*-score showed more significant decrease in the quality of the data when compared with the Loess-fit data under hit rates above 20%. *B*-score performed reliably only below a 20% hit rate, which is a much more narrow range than for the loess-fit. Thus, *B*-score normalization may lead to a high number of false positives for experiments with a hit rate above 20%. The Loess-fit method generated data of higher reliability (Fig. 5a and b), close to or even higher than the recommended thresholds of QC metrics.

Using the plate layout with controls at the edge resulted in the lowest possible post-normalization QC scores for both *B*-score and Loess normalized data. The layout with controls scattered led to

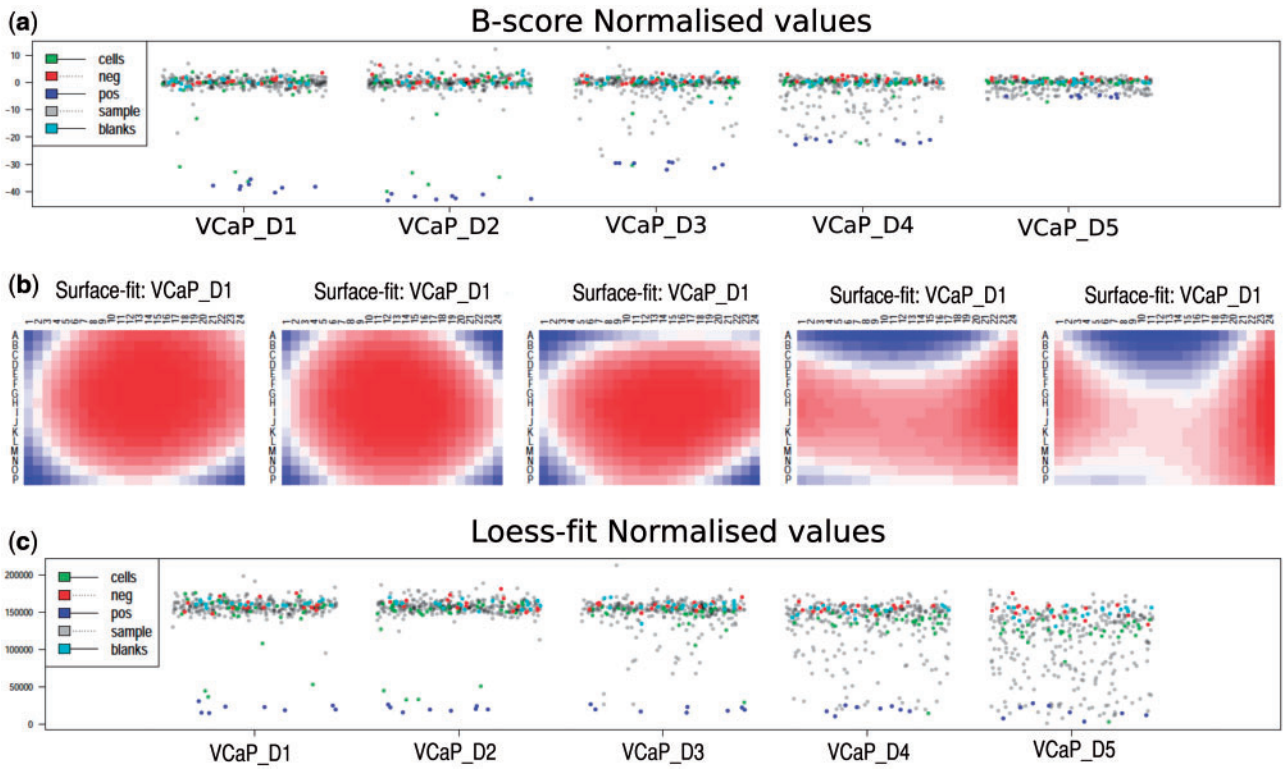

Fig. 6. Post-normalization QC assessment. Data quality of normalized data is inspected using scatter plots (**a**) and (**c**). The surface-fit images (**b**) are generated from the raw data so that we can detect areas with uneven concentration of hits (red) or no hits (blue) emanating from with-in plate effects. The plate scatters are organized in ascending order of concentration (D1–D5) as shown on the *x*-axis labels. The positive (blue) and negative (red) controls represent maximum cell killing effect and no cell killing effect respectively. Consequently, the over scaled data in (a) by the *B*-score in the fifth plate D5 with a high hit rate, means reduced quality of data whereas (c) represents good quality post-normalization data

better QC scores for data based on both normalization methods under low hit rate conditions. Therefore, the two simulation studies revealed that, in a high throughput experiment with a hit rate less than 20%, a combination of the Loess-fit normalization method and the layout with controls scattered performed better than the *B*-score.

3.3 Application of *B*-score and Loess methods to real drug testing experimental data

To confirm the conclusions based on our simulation study, we compared normalizations with the *B*-score and Loess methods using real drug testing data with the VCaP and LAPC4 screens (Table 1). Our results confirmed generation of poor QC scores using *B*-score normalization for the fourth and fifth plates containing drugs applied at high concentrations. It was evident that normalization of the data using the *B*-score method altered signal intensities for controls as the number of hits increased on plates (Fig. 6a). Loess-fit normalization was applied on the same data, resulting in a surface fit exposing areas with uneven distribution of low and high values. The areas showing systematic high or low values in adjacent wells could be due to evaporation induced edge effects or uneven cell seeding. Figure 6b illustrates the Loess-fit heatmaps showing some plates with edge effects. In an ideal screen, the fitted surface will identify no spots adjacent to each other with systematically high or low signal intensities (hills and valleys), thus leading to a zero surface flat plane. When systematic within-plate errors do exist in some positions on the plate, the residuals for the fitted surface around such wells will be high and need to be corrected. Compared with the *B*-score method, the scatter plots based on Loess-fit approach retained a good dynamic range between control samples placed on the plates (Fig. 6c). There was no major alteration of control well signals for plates containing drugs applied at high concentration (Fig. 6a). Loess-fit approach offers an improved way of normalizing HTS data under high hit rate scenarios compared with previous. Next, we assessed the quality of data for replicate experiments after preforming normalization using the *Z*'-factor and SSMD scores (Table 1).

The QC metrics always flagged the fifth plate as being of poor quality, which was not correct based on the QC scores observed from using raw data for the same screens. Given the inconsistency between pre and post-normalization QC results for the fifth plate, we concluded that normalization of any screen with a high hit rate leads to an increase in false-positive hits. We observed that the hit rate is the most critical factor that causes normalizations to fail. From the real screening data, we also observed that no normalization method could correct for strong within-plate effects and thereby improve QC scores as observed in the VCaP replicate 2 experiments. Consequently, when a drug testing experiment has poor QC metrics such as Z'-factor < 0.5 and SSMD < 6 coupled with a high hit rate (>20%), it is worth repeating the experiment rather than performing any normalization on the data.

4 Conclusions

We have shown the importance of QC metrics and per-plate data visualization in identifying systematic errors in HTS experiments with a high hit rate. Our results show that a scattered layout is superior over the layout based on placing controls in the first and last columns. Careful assessment of the impact of normalization is needed particularly when the original assumptions on low hit rates

are violated. Furthermore, our results show that the Loess-fit method performs better than the *B*-score method for most experiment scenarios, especially when the hit rate is below 20%. We observed that the quality of data from experiments with plates containing a hit rate above 20% will be severely compromised by any normalization method and could lead to a large number of false-positive hits. Most normalization methods are designed with the assumption that hits are few in number and are sparsely distributed across the entire plate (Murie *et al*., 2013). However, in the case of drug testing with multiple doses for each drug, plates are often made from serial dilutions and thereby the plate with the highest concentration of drugs will contain rows and columns with many hits. Common normalization methods for correcting row and column effects will adjust these wells on the plate since they will be detected as within plate artifacts. Our data point to *B*-score's main defect being the assumption of a low hit rate on every row and column for all plates. To calculate the *B*-score, the resulting residuals within each plate are divided by their median absolute deviation as a standardization step. The *B*-score calculation for a given position is therefore impacted by a high number of hits on the row and column as the median polishing is performed. Also, *B*-score does not weigh local effects identified by distribution of the hits that can be visualized by surface fitting.

We postulate that the good performance of the Loess-fit is because it is based on fitting a local distribution surface rather than adjusting effects based on row and column signal intensities. The strong advantage of the local smooth surface fitting procedure is that it discovers areas on a plate where signal intensities are systematically higher or lower that may be concentrated in adjacent wells on a plate. We have similar observations with the developers of ChemBank repository (Seiler *et al*., 2008) on the challenges of using the *B*-score on high hit rate screens.

In summary, tailored approaches are needed for high hit-rate HTS experiments as well as for drug testing where dose–response curves are acquired. HTS studies and automated HTS data analysis software using any normalization or scaling scheme should include an option to inspect the quality of the raw data before normalization and report the detailed description of the data processing procedures used in the analyses. By reporting the pre- and post-normalization QC results for all HTS drug-testing experiments, the quality and reproducibility of all HTS data would be improved.

Acknowledgements

This work was supported by Helsinki Biomedical Graduate School, Institute for Molecular Medicine Finland, Academy of Finland, Sigrid Juselius Foundation and the Cancer Organizations of Finland. We thank Henri Sara for his contribution to the development of the Loess-fit HTS method and also do thank Dr Pekka Kohonen for helping us with HTS data analysis interpretations and method testing. Senior laboratory technician Mariliina Arjamaa is thanked for excellent technical assistance in performing drug testing on VCaP and LAPC4 cells.

Funding

The research leading to these results has been supported by the European Union's Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 258068; EU-FP7-Systems Microscopy NoE.

Conflict of Interest: none declared.

References

Barretina,J. *et al*. (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.

Birmingham,A. *et al*. (2009) Statistical methods for analysis of high-throughput RNA interference screens. *Nat. Methods*, **6**, 569–575.

Boutros,M. *et al*. (2006) Analysis of cell-based RNAi screens. *Genome Biol.*, **7**, R66.

Brideau,C. *et al*. (2003) Improved statistical methods for hit selection in high-throughput screening. *J. Biomol. Screen.*, **8**, 634–647.

Crystal,A.S. *et al*. (2014) Patient-derived models of acquired resistance can identify effective drug combinations for cancer. *Science*, **346**, 1480–1486.

Dragiev,P. *et al*. (2011) Systematic error detection in experimental high-throughput screening. *BMC Bioinformatics*, **12**, 25.

Dragiev,P. *et al*. (2012) Two effective methods for correcting experimental high-throughput screening data. *Bioinformatics*, **28**, 1775–1782.

Gao,D. *et al*. (2014) Organoid cultures derived from patients with advanced prostate cancer. *Cell*, **159**, 176–187.

Garnett,M.J. *et al*. (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, **483**, 570–575.

Hatzis,C. *et al*. (2014) Enhancing reproducibility in cancer drug screening: how do we move forward?. *Cancer Res.*, **74**, 4016–4023.

Kafadar,K. (2003) John Tukey and robustness. *Stat. Sci.*, **18**, 319–331.

Kwan,P. and Birmingham,A. (2010) NoiseMaker: simulated screens for statistical assessment. *Bioinformatics*, **26**, 2484–2485.

Lin,L.I. (1989) A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, **45**, 255–268.

Liu,X. *et al*. (2013) HitPick: a web server for hit identification and target prediction of chemical screenings. *Bioinformatics*, **29**, 1910–1912.

Makarenkov,V. *et al*. (2007) An efficient method for the detection and elimination of systematic error in high-throughput screening. *Bioinformatics*, **23**, 1648–1657.

Malo,N. *et al*. (2006) Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol.*, **24**, 167–175.

Mangat,C.S. *et al*. (2014) Rank ordering plate data facilitates data visualization and normalization in high-throughput screening. *J. Biomol. Screen.*, **19**, 1314–1320.

Murie,C. *et al*. (2013) Control-plate regression (CPR) normalization for high-throughput screens with many active features. *J. Biomol. Screen.*, **19**, 661–671.

Pemovska,T. *et al*. (2013) Individualized systems medicine strategy to tailor treatments for patients with chemorefractory acute myeloid leukemia. *Cancer Discov.*, **3**, 1416–1429.

Rieber,N. *et al*. (2009) RNAither, an automated pipeline for the statistical analysis of high-throughput RNAi screens. *Bioinformatics*, **25**, 678–679.

Risso,D. *et al*. (2009) A modified LOESS normalization applied to microRNA arrays: a comparative evaluation. *Bioinformatics*, **25**, 2685–2691.

Seiler,K.P. *et al*. (2008) ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.*, **36**, D351–9.

Shapiro,S.S. and Wilk,M. (1965) An analysis of variance test for normality (complete samples). *Biometrika*, **52**, 591–611.

Shoemaker,R.H. (2006) The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer*, **6**, 813–823.

Tyner,J.W. *et al*. (2013) Kinase pathway dependence in primary human leukemias determined by rapid inhibitor screening. *Cancer Res.*, **73**, 285–296.

Xiong,H. *et al*. (2008) Using generalized procrustes analysis (GPA) for normalization of cDNA microarray data. *BMC Bioinformatics*, **9**, 25.

Yadav,B. *et al*. (2014) Quantitative scoring of differential drug sensitivity for individually optimized anticancer therapies. *Sci. Rep.*, **4**, 5193.

Yang,W. *et al*. (2013) Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.*, **41**, D955–D961.

Zhang,J.H. *et al*. (1999) A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J. Biomol. Screen.*, **4**, 67–73.

Zhang,X.D. (2007) A pair of new statistical parameters for quality control in RNA interference high-throughput screening assays. *Genomics*, **89**, 552–561.

Zhang,X.D. (2011) Illustration of SSMD, z score, SSMD*, z* score, and t statistic for hit selection in RNAi high-throughput screens. *J. Biomol. Screen.*, **16**, 775–785.

Zhang,X.D. *et al*. (2007) The use of strictly standardized mean difference for hit selection in primary RNA interference high-throughput screening experiments. *J. Biomol. Screen.*, **12**, 497–509.